DOCUMENT RESUME

ED 434 932 TM 030 184

AUTHOR Tanguma, Jesus
TITLE Variable Deletion Strategies in Canonical Correlation Analysis.
PUB DATE 1999-11-17
NOTE 23p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Point Clear, AL, November 1999).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Functions (Mathematics); *Multivariate Analysis; Research Methodology; Researchers
IDENTIFIERS *Weighting (Statistical)

ABSTRACT

This paper presents three variable deletion strategies in canonical correlation analysis. All three strategies are illustrated by examples. The first strategy uses the canonical communality (h2) coefficients of the three functions to decide which variable to delete. The second function also uses the canonical communality coefficients, but only after deleting the least contributing function. The third strategy used weighted canonical communality coefficients on all three functions to decide which variable to delete. All three strategies attempt to provide the researcher with a more parsimonious canonical solution. (Contains 11 tables and 4 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 434 932

Running head: VARIABLE DELETION STRATEGIES

Variable Deletion Strategies in Canonical Correlation Analysis

Jesus Tanguma

Texas A & M University

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Jesus Tanguma

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the Mid-South Educational Research Association, Point Clear, AL, November 17, 1999

TM030184

BEST COPY AVAILABLE

Abstract

The present paper presents three variable deletion strategies in canonical correlation analysis. All three strategies are illustrated by means of an example. The first strategy uses the canonical communality (h^2) coefficients of the three functions to decide which variable to delete. The second function also uses the canonical communality (h^2) coefficients but only after deleting the least contributing function. The third strategy uses weighted canonical communality (h^2) coefficients on all three functions to decide which variable to delete. All three strategies strive to provide the researcher with a more parsimonious canonical solution.

Variable Deletion Strategies in Canonical Correlation Analysis

Researchers, in an effort to make the results of a study more parsimonious and generalizable, often wish to delete predictor variables from a study. In other words, the researchers feel that more is not always better, but sometimes it may be worse. Simply put, by reducing the number of predictor variables, a more parsimonious canonical solution may be obtained. Consequently, the canonical solution may be more likely to be true and replicable (Thompson, 1984).

In regression analysis, the most frequently used variable deletion methods are the so-called stepwise methods. However, as Thompson (1996) has repeatedly stated, these methods are inherently flawed and should not be used for this or other purposes. In canonical correlation analysis, methods have been developed to assist the researcher in deciding which variables or functions to delete.

The purpose of the present paper is to discuss three variable deletion strategies for use in canonical correlation analysis. To do so, a data set originally analyzed by Holzinger and Swineford (1939) will be used.

First Strategy

The first variable deletion strategy uses the canonical communality (h^2) coefficients to decide which variable to drop first. As Thompson (1991) pointed out, the canonical communality coefficients are "indicative of how much of the variance of an observed variable is contained within the set of synthetic variables" (p. 86). The canonical communality coefficients are obtained by adding the squared structure coefficient for each canonical function. For example, the canonical communality coefficient for T20, see Table 1, is obtained by adding

0.5944	the squared structure coefficient for function I
0.1218	the squared structure coefficient for function II
0.2841	the squared structure coefficient for function III
1.00	

Thus, indicating that the three synthetic variables can reproduce 100% of the variance of T20.

Another important factor in determining whether or not to drop a given variable is to look at the change of the squared canonical correlation coefficient (R^2_c). The squared canonical correlation coefficient (R^2_c) explains how much each function is contributing to the overall canonical solution. If there is little change in the R^2_c after dropping a variable, the resulting canonical solution would be more parsimonious.

Insert Table 1 About Here

Since T17 had the lowest h^2 (15.50%), it was the first variable dropped. Table 2 presents the complete canonical solution after dropping T17. Notice that the squared canonical correlation coefficient (R^2_c) for Function I did not change. The change in R^2_c

for the second and third functions was very small (0.10% for each). Also, there were very small changes in the function, structure, and communality coefficients.

Insert Table 2 About Here

However, the canonical communality coefficient (h^2) for variable T18 (50.71%) was considerably lower than the others. Therefore, T18 was the next variable dropped. Table 3 presents the complete canonical solution after dropping this variable. Notice the small changes in R^2_c for each function (i.e., 0.40 for Function I, 0.20 for Function II, and 0.10 for Function III). There were also very small changes in the function, structure, and communality coefficients. Thus, dropping variable T18 resulted in a more parsimonious canonical solution.

Insert Table 3 About Here

Second Strategy

A second strategy in variable deletion involves looking at the contribution of the individual functions. As Table 1 indicates, the third function is contributing the least to the overall canonical solution ($h^2 = 0.60\%$). Consequently, the researcher might want to drop this function. Table 4 presents the complete canonical solution after dropping the third function. Notice that, again, T17 has the lowest h^2. Consequently, T17 is the first variable to be dropped. Notice that the squared canonical correlation coefficient (R^2_c) for Function I did not change. The change in R^2_c for Function II was very small (0.10%). Also, there were very small changes in the function, structure, and communality

coefficients. Therefore, dropping variable T17 resulted in a more parsimonious canonical solution.

Insert Tables 4 and 5 About Here

However, the canonical communality coefficient (h^2) for variable T14 (19.80%) was considerably lower than the others. Therefore, T14 was the next variable dropped. Table 6 presents the complete canonical solution after dropping this variable. Notice the small changes in R^2_c for each function (i.e., 0.10 for Function I and 0.20 for Function II). There were also very small changes in the function, structure, and communality coefficients. Thus, dropping variable T14 resulted in a more parsimonious canonical solution.

Insert Table 6 About Here

Although a more parsimonious solution had been obtained by dropping T17 and T14, the canonical communality coefficient for T18 was remarkably lower than the others (h^2 =26.40%). Therefore, T18 was dropped. Table 7 presents the complete canonical solution after dropping variable T18. The remaining h^2s were very homogeneous in their numerical values.

Insert Table 7 About Here

Third Strategy

A third variable deletion strategy requires computing a weighted h^2. This weighted h^2 reflects the variable's contribution to the function as well as the function's

contribution to the complete canonical solution. Thus, the weighted h^2 provides a clearer picture as to each variable's total contribution to the complete canonical solution.

The weighted h^2 for each measured variable is found by multiplying the squared structure coefficient for each function times the R^2_c for each function and then summing up all the products. For example, the weighted h^2 for T20 is

(0.5644*0.2790 + 0.1218*0.03 + 0.2841*0.006) = 0.1712 or 17.12%.

Table 8 presents the complete canonical solution using weighted h^2s. Notice that T17 has the lowest weighted h^2. Consequently, T17 was the first variable dropped. Table 9 presents the complete canonical solution after dropping T17. Notice that the R^2_c for Function I did not change. The changes in R^2_c for the second and third functions were very small (0.10% for each). Also, there were very small changes in the function, structure, and weighted communality coefficients. Therefore, dropping variable T17 resulted in a more parsimonious canonical solution.

Insert Tables 8 and 9 About Here

However, the weighted canonical communality coefficient (R^2_c) for variable T14 (5.62%) was lower than all the others. Therefore, T14 was the next variable dropped. Table 10 presents the complete canonical solution after dropping variable T14. The changes in R^2_c for the first, second, and third functions were 0.10%, 0.20%, and 0.30%, respectively. There were also very small changes in the function, structure, and weighted communality coefficients. Thus, dropping variable T14 resulted in a more parsimonious canonical solution.

Insert Table 10 About Here

Although the weighted communality coefficients (h^2) for variables T15 (6.45%) and T18 (6.61%) were remarkably lower than all the others, these variables were not dropped from the final canonical solution. However, the effect of dropping each variable was investigated. For example, when variable T15 was dropped, the R^2_c for Function III remained the same. But, the R^2_c for Function I changed from 27.80% to 26.70%. Also, there was a big drop in R^2_c for Function II. That is, the R^2_c for Function II changed from 2.70% to 1.0%. Therefore, the researcher decided not to drop this (T15) variable. Similarly, when variable T18 was dropped, the squared canonical correlation coefficient (R^2_c) for the third function was computed to be zero. Consequently, the researcher decided not to drop this (T18) variable.

Table 11 presents a summary of the changes in R^2_c for each variable deletion strategy as well as the effects of deleting a particular variable.

Insert Table 11 About Here

Conclusion

Three variable deletion strategies were presented and applied to nine variables (i.e., three in one set, and six in the other set) from the Holzinger and Swineford (1939) data. All three strategies strive to provide the researcher with a more parsimonious canonical solution. In other words, all three strategies try to reproduce approximately the same amount of variance with a reduced (smaller) variable set than with the original (larger) variable set. The smaller set is a more parsimonious solution, and consequently more likely to be true and replicable (Thompson, 1984).

Glossary

Canonical Correlation Coefficient (R^2_c) The Pearson product-moment correlation between the two sets of synthetic variable scores computed for a given canonical function.

Communality Coefficient (h^2) The proportion or percentage of variance in a measured variable that is useful in defining the canonical solution.

Function The set (in some analyses called equation or factor) of weights (e.g., regression beta weights, factor pattern coefficients, canonical function coefficients) applied to the measured variables to yield scores on synthetic variables (e.g., regression YHAT scores, factor scores, canonical scores).

Function Coefficient The multiplicative constant or weight applied to a given measured variable as part of the calculation of scores on synthetic variables.

Redundancy Coefficient (Rd) A canonical coefficient in a squared metric that is not multivariate, and which is useful in CCA only in very unusual cases where a "g" function with perfect effect size is expected.

Structure Coefficient (r_s) The Pearson product-moment correlation between scores on a given measured variable and the synthetic variable scores on a given function for the variable set to which the measured variable belongs.

Synthetic/Latent Variable Estimates of latent constructs computed by applying weights to the measured variables (e.g., regression YHAT scores, factor scores, canonical function scores).

Weight The multiplicative constants (e.g., regression beta weights, factor pattern coefficients, canonical function coefficients) applied to the measured variables to

yield scores on synthetic variables (e.g., regression YHAT scores, factor scores, canonical function scores).

References

Thompson, B. (1984). Canonical correlation analysis: Uses and Interpretation. Newbury Park, CA: SAGE.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. Measurement and Evaluation in Counseling and Development, 24(2) 80-95.

Thompson, B. (1996). Canonical correlation analysis. In L. Grimm & P. Yarnold (Eds). Reading and understanding multivariate statistics, 2. Washington, DC: American Psychological Association.

Table 1

Complete canonical solution

	Function I			Function II			Function III			
	Funct	Struct	Struct2	Funct	Struct	Struct2	Funct	Struct	Struct2	H2
T20	-0.434	-0.771	59.44%	0.652	0.349	12.18%	0.821	0.533	28.41%	100.00%
T21	-0.525	-0.82	72.25%	0.294	0.15	2.25%	-0.952	-0.553	30.58%	78.08%
T22	-0.334	-0.703	49.42%	-1.057	-0.689	47.47%	0.209	0.175	3.07%	99.96%
Adequacy			60.37%			20.63%			20.69%	
Rd			16.84%			0.62%			0.12%	
Rc2			27.90%			3.00%			0.60%	
Rd			9.14%			2.42%			0.57%	
Adequacy			32.77%			80.76%			95.73%	
T14	0.29	-0.437	19.20%	-0.356	-0.081	0.66%	0.878	0.717	51.41%	71.27%
T15	-0.11	-0.415	17.22%	0.925	0.718	51.55%	0.189	0.433	18.75%	87.52%
T16	-0.625	-0.831	69.06%	0.287	0.256	6.55%	-0.437	-0.063	0.39%	76.00%
T17	-0.034	-0.374	13.99%	-0.059	-0.034	0.12%	-0.263	0.118	1.39%	15.50%
T18	-0.116	-0.478	22.85%	-0.327	-0.204	4.16%	0.542	0.479	22.94%	49.95%
T19	-0.514	-0.737	54.32%	-0.392	-0.421	17.72%	-0.352	-0.092	0.85%	72.89%

Table 2
Variable t17 has been dropped.

	Function I			Function II			Function III			
	Funct	Struct	Struct2	Funct	Struct	Struct2	Funct	Struct	Struct2	H2
T20	-0.435	-0.772	59.59%	0.646	0.345	11.91%	0.825	0.534	28.52%	100.00%
T21	-0.524	-0.819	67.08%	0.30	0.153	2.34%	-0.950	-0.552	30.47%	99.89%
T22	-0.334	-0.703	49.42%	-1.058	-0.690	47.61%	0.202	0.170	2.89%	99.92%
Adequacy			58.69%			20.62			20.63%	
Rd			16.65%			0.61%			0.10%	
Rc2			27.90%			2.90%			0.50%	
Adequacy			12.75%			0.59%			0.12%	
Rd			45.70%			20.19%			24.86%	
T14	0.022	-0.437	19.09%	-0.371	-0.084	0.71%	0.839	0.733	53.73%	73.53%
T15	-0.114	-0.416	17.31%	0.919	0.717	51.41%	0.176	0.453	20.52%	89.24%
T16	-0.627	-0.832	69.22%	0.284	0.255	6.50%	-0.466	-0.067	0.45%	76.17%
T18	-0.125	-0.478	22.85%	-0.344	-0.206	4.24%	0.481	0.486	23.62%	50.71%
T19	-0.516	-0.737	54.32%	-0.395	-0.423	17.89%	-0.384	-0.106	1.12%	73.33%

Table 3
Variables t17 and t18 have been dropped.

	Function I			Function II			Function III			
	Funct	Struct	Struct2	Funct	Struct	Struct2	Funct	Struct	Struct2	H2
T20	-0.438	-0.772	59.59%	0.702	0.376	14.14%	0.777	0.513	26.13%	99.86%
T21	-0.534	-0.825	68.06%	0.221	0.101	1.02%	-0.966	-0.556	30.91%	99.99%
T22	-0.319	-0.693	48.02%	-1.044	-0.684	46.79%	0.284	0.227	5.15%	99.96%
Adequacy			58.56%			20.65%			20.73%	
Rd			16.10%			0.56%			0.01%	
Rc2			27.50%			2.70%			0.40%	
Adequacy			11.12%			0.57%			0.01%	
Rd			40.44%			21.21%			22.90%	
T14	0.021	-0.439	19.27%	-0.37	-0.08	0.64%	1.037	0.87	75.69%	95.60%
T15	-0.147	-0.421	17.72%	0.883	0.75	56.25%	0.163	0.392	15.37%	89.34%
T16	-0.644	-0.838	70.22%	0.226	0.232	5.38%	-0.487	-0.073	0.53%	76.13%
T19	-0.55	-0.74	54.76%	-0.539	-0.475	22.56%	-0.156	0.013	0.01%	77.33%

Table 4
Function #3 has been deleted.

	Function I			Function II			
	Funct	Struct	Struct2	Funct	Struct	Struct2	H2
T20	-0.434	-0.771	59.44%	0.652	0.349	12.18%	71.62%
T21	-0.525	-0.82	72.25%	0.294	0.15	2.25%	74.50%
T22	-0.334	-0.703	49.42%	-1.057	-0.689	47.47%	96.89%
Adequacy			60.37%			20.63%	
Rd			16.84%			0.62%	
Rc2			27.90%			3.00%	
Rd			9.14%			2.42%	
Adequacy			32.77%			80.76%	
T14	0.29	-0.437	19.20%	-0.356	-0.081	0.66%	19.86%
T15	-0.11	-0.415	17.22%	0.925	0.718	51.55%	68.77%
T16	-0.625	-0.831	69.06%	0.287	0.256	6.55%	75.61%
T17	-0.034	-0.374	13.99%	-0.059	-0.034	0.12%	14.11%
T18	-0.116	-0.478	22.85%	-0.327	-0.204	4.16%	27.01%
T19	-0.514	-0.737	54.32%	-0.392	-0.421	17.72%	72.04%

Table 5
Function #3 has been deleted. Variable t17 has been dropped.

	Function I			Function II			
	Funct	Struct	Struct2	Funct	Struct	Struct2	H2
T20	-0.435	-0.772	59.59%	0.646	0.345	11.90%	71.49%
T21	-0.524	-0.819	67.08%	0.3	0.153	2.34%	69.42%
T22	-0.334	-0.703	49.42%	-1.058	-0.69	47.61%	97.03%
Adequacy			58.70%			20.62%	
Rd			16.38%			0.59%	
Rc2			27.90%			2.90%	
Rd			10.20%			0.46%	
Adequacy			36.56%			16.15%	
T14	0.022	-0.437	19.09%	-0.371	-0.084	0.71%	19.80%
T15	-0.114	-0.416	17.31%	0.919	0.717	51.41%	68.72%
T16	-0.627	-0.832	69.22%	0.284	0.255	6.51%	75.73%
T18	-0.125	-0.478	22.85%	-0.344	-0.206	4.24%	27.09%
T19	-0.516	-0.737	54.32%	-0.395	-0.423	17.89%	72.21%

Table 6
Function #3 has been deleted.
Variables t17 and t14 have been dropped because low h2.

	Function I			Function II			
	Funct	Struct	Struct2	Funct	Struct	Struct2	H2
T20	-0.436	-0.772	59.59%	0.74	0.407	16.56%	76.15%
T21	-0.521	-0.818	66.91%	0.187	0.089	0.79%	67.70%
T22	-0.337	-0.705	49.70%	-1.027	-0.664	44.09%	93.79%
Adequacy			58.73%			20.48%	
Rd			16.33%			0.55%	
Rc2			27.80%			2.70%	
Rd			11.39%			0.61%	
Adequacy			40.96%			22.45%	
T15	-0.108	-0.415	17.22%	0.867	0.773	59.75%	76.97%
T16	-0.622	-0.832	69.22%	0.194	0.267	7.12%	76.34%
T18	-0.125	-0.479	22.94%	-0.341	-0.186	3.46%	26.40%
T19	-0.513	-0.738	54.46%	-0.486	-0.441	19.45%	73.41%

Table 7
Function #3 has been deleted.
Variables t17, t14, and t18 have been dropped because of low h2.

	Function I			Function II			
	Funct	Struct	Struct2	Funct	Struct	Struct2	H2
T20	-0.438	-0.772	59.59%	0.795	0.438	19.18%	78.77%
T21	-0.532	-0.824	67.90%	0.098	0.031	0.01%	67.91%
T22	-0.322	-0.695	48.30%	-0.999	-0.649	42.12%	90.42%
Adequacy			58.60%			20.44%	
Rd			16.12%			0.49%	
Rc2			27.50%			2.40%	
Rd			13.09%			0.75%	
Adequacy			47.62%			31.28	
T15	-0.141	-0.421	17.72%	0.828	0.801	64.16%	81.88%
T16	-0.639	-0.838	70.22%	0.134	0.242	5.86%	76.08%
T19	-0.547	-0.741	54.91%	-0.622	-0.488	23.81%	78.72%

Table 8
Complete canonical solution with weighted h2.

	Function I			Function II			Function III			
	Funct	Struct	Struct2	Funct	Struct	Struct2	Funct	Struct	Struct2	weighted h2
T20	-0.434	-0.771	59.44%	0.652	0.349	12.18%	0.821	0.533	28.41%	17.12%
T21	-0.525	-0.82	72.25%	0.294	0.15	2.25%	-0.952	-0.553	30.58%	20.41%
T22	-0.334	-0.703	49.42%	-1.057	-0.689	47.47%	0.209	0.175	3.07%	15.23%
Adequacy			60.37%			20.63%			20.69%	
Rd			16.84%			0.62%			0.12%	
Rc2			27.90%			3.00%			0.60%	
Rd			9.14%			2.42%			0.57%	
Adequacy			32.77%			80.76%			95.73%	
T14	0.29	-0.437	19.20%	-0.356	-0.081	0.66%	0.878	0.717	51.41%	5.69%
T15	-0.11	-0.415	17.22%	0.925	0.718	51.55%	0.189	0.433	18.75%	6.46%
T16	-0.625	-0.831	69.06%	0.287	0.256	6.55%	-0.437	-0.063	0.39%	19.47%
T17	-0.034	-0.374	13.99%	-0.059	-0.034	0.12%	-0.263	0.118	1.39%	3.92%
T18	-0.116	-0.478	22.85%	-0.327	-0.204	4.16%	0.542	0.479	22.94%	6.64%
T19	-0.514	-0.737	54.32%	-0.392	-0.421	17.72%	-0.352	-0.092	0.85%	15.69%

Table 9
Canonical solution with weighted h2. Variable t17 has been dropped.

	Function I			Function II			Function III			
	Funct	Struct	Struct2	Funct	Struct	Struct2	Funct	Struct	Struct2	weighted h2
T20	-0.435	-0.772	59.59%	0.646	0.345	11.91%	0.825	0.534	28.52%	17.11%
T21	-0.524	-0.819	67.08%	0.30	0.153	2.34%	-0.950	-0.552	30.47%	18.94%
T22	-0.334	-0.703	49.42%	-1.058	-0.690	47.61%	0.202	0.170	2.89%	15.18%
Adequacy			58.69%			20.62			20.63%	
Rd			16.65%			0.61%			0.10%	
Rc2			27.90%			2.90%			0.50%	
Adequacy			12.75%			0.59%			0.12%	
Rd			45.70%			20.19%			24.86%	
T14	0.022	-0.437	19.09%	-0.371	-0.084	0.71%	0.839	0.733	53.73%	5.62%
T15	-0.114	-0.416	17.31%	0.919	0.717	51.41%	0.176	0.453	20.52%	6.42%
T16	-0.627	-0.832	69.22%	0.284	0.255	6.50%	-0.466	-0.067	0.45%	19.50%
T18	-0.125	-0.478	22.85%	-0.344	-0.206	4.24%	0.481	0.486	23.62%	6.62%
T19	-0.516	-0.737	54.32%	-0.395	-0.423	17.89%	-0.384	-0.106	1.12%	15.68%

Table 10
Canonical solution with weighted h2. Variables t17 and t14 have been dropped.

	Function I			Function II			Function III			
	Funct	Struct	Struct2	Funct	Struct	Struct2	Funct	Struct	Struct2	weighted h2
T20	-0.436	-0.772	59.59%	0.74	0.407	16.56%	0.742	0.488	23.81%	17.01%
T21	-0.521	-0.818	66.91%	0.187	0.089	0.79%	-0.98	-0.568	32.26%	18.69%
T22	-0.337	-0.705	49.70%	-1.027	-0.664	44.09%	0.325	0.248	6.15%	15.02%
Adequacy			58.73%			20.48%			20.74	
Rd			16.33%			0.55%			0.01%	
Rc2			27.80%			2.70%			0.20%	
Rd			11.39%			0.61%			0.01%	
Adequacy			40.96%			22.45%			22.09%	
T15	-0.108	-0.415	17.22%	0.867	0.773	59.75%	0.293	0.379	14.36%	6.43%
T16	-0.622	-0.832	69.22%	0.194	0.267	7.12%	-0.512	-0.233	5.43%	19.45%
T18	-0.125	-0.479	22.94%	-0.341	-0.186	3.46%	0.929	0.828	68.56%	6.61%
T19	-0.513	-0.738	54.46%	-0.486	-0.441	19.45%	-0.19	0.002	0.01%	15.67%

Table 11
Summary of changes in R^2_c

	Function I	Function II	Function III
	R^2_c	R^2_c	R^2_c
Variables deleted			
	27.90%	3.00%	0.60%
T17	27.90%	2.90%	0.50%
T17, T18	27.50%	2.70%	0.40%
	27.90%	3.00%	
T17	27.90%	2.90%	
T17,T14	27.80%	2.70%	
T17, T14, T18	27.50%	2.40%	
	27.90%	3.00%	0.60%
T17	27.90%	2.90%	0.50%
T17,T14	27.80%	2.70%	0.20%